Tratamiento lingüístico de las preguntas en español en los sistemas de búsqueda de respuestas

Por **María-Dolores Olvera-Lobo** y **Nicolás Robinson-García**

Resumen: *Se propone un procedimiento para el tratamiento lingüístico de las preguntas en español como paso previo a su clasificación en los sistemas de búsqueda de respuestas. Se mencionan los principales tipos de sistemas de búsqueda de respuestas y su arquitectura básica. Se revisan las principales taxonomías utilizadas hasta el momento para la clasificación de preguntas y las distintas perspectivas desde las que se enfocan. Finalmente, se presentan las etapas de análisis lingüístico a las que ha de someterse el texto de las preguntas en estos sistemas para facilitar la localización de las respuestas adecuadas.*

Palabras clave: *Sistemas de búsqueda de respuestas, Clasificación de preguntas, Procesamiento del lenguaje natural, Análisis lingüístico de las preguntas, Lengua española.*

Title: Linguistic treatment of questions in Spanish for question classification in question answering systems.

Abstract: *We propose a procedure for the linguistic treatment of Spanish questions as a step prior to their classification in question answering systems. The main types of question answering systems and their basic architecture are described. We review the principal question classification taxonomies used to date and the different fields from which they have been derived. Finally, we present the stages of linguistic analysis that the text of questions in question answering systems should be subject to in order to facilitate the location of appropriate answers.*

Keywords: *Question answering systems, Question classification, Natural language processing, Linguistic analysis of questions, Spanish language.*

Olvera-Lobo, María-Dolores; Robinson-García, Nicolás. "Tratamiento lingüístico de las preguntas en español para su clasificación en los sistemas de búsqueda de respuestas". *El profesional de la información*, 2009, marzo-abril, v. 18, n. 2, pp. 180-187.

DOI: 10.3145/epi.2009.mar.08

María-Dolores Olvera-Lobo *es doctora en documentación, profesora titular del Depto. de Biblioteconomía y Documentación, y docente en la Fac. de Comunicación y Documentación y en la Fac. de Traducción e Interpretación de la Univ. de Granada. Pertenece al grupo Scimago (Unidad Asociada del Csic). Participa en varios proyectos de investigación y es autora de libros y artículos especializados.*

Nicolás Robinson-García *es diplomado en biblioteconomía y documentación por la Universidad de Granada. Ha colaborado en servicios de documentación y en la sección de cultura en algunos medios de comunicación escritos. En la actualidad sus intereses se centran en la investigación sobre sistemas avanzados de recuperación de información.*

Introducción

A lo largo de la última década se ha realizado una serie de esfuerzos por desarrollar la que se ha bautizado como Web semántica (**Berners-Lee** et al., 2001). Desde el punto de vista de la recuperación de la información, el acceso a los contenidos de la Web semántica puede favorecerse con el uso del lenguaje natural, ya que a los usuarios les resulta más cómodo y ágil usar su forma habitual de expresión (**Vállez**, 2007). La disciplina denominada procesamiento del lenguaje natural (PLN) resulta especialmente útil en este contexto puesto que se encarga de estudiar los problemas de la generación y la comprensión automática del lenguaje natural. El PLN también se centra en diseñar sistemas y mecanismos eficaces que permitan la comunicación entre personas y máquinas.

En el acceso a la información en internet ya no se trata únicamente de recuperar documentos, sino de extraer conocimiento de los mismos y de elaborar respuestas que satisfagan las necesidades de los usuarios. Para ello es necesaria la implementación de sistemas de recuperación de información capaces de procesar el lenguaje natural y de "comprender" tanto las consultas que plantea el usuario como la información almacena-

Artículo recibido el 10-10-08
Aceptación definitiva: 30-12-08

da en su base de datos. Entre los muchos tipos de sistemas que responden a esta filosofía podemos citar los de búsqueda de respuestas (en inglés, *question answering systems* o *QA systems*).

El principal objetivo de este trabajo es plantear una propuesta para la adaptación de las clásicas etapas de análisis lingüístico (fonético, morfológico, léxico, sintáctico, semántico, discursivo y pragmático) al tratamiento de las preguntas en español con el fin de facilitar la clasificación de las mismas en los sistemas de búsqueda de respuestas. Se describe la arquitectura y funcionamiento general de estos sistemas y se hace hincapié en las etapas del análisis y clasificación de las preguntas de usuario. A continuación, se revisan las principales taxonomías que se han sugerido para la clasificación de las preguntas, las diferentes perspectivas empleadas, y la utilidad que se les ha proporcionado hasta el momento. Por último, se detallan las etapas del tratamiento lingüístico de las preguntas con el objeto de establecer un punto de partida para futuras investigaciones en este ámbito.

Tipos de sistemas de búsqueda de respuestas

Estos sistemas persiguen ofrecer una solución a preguntas concretas de los usuarios formuladas en lenguaje natural. Las preguntas que principalmente gestionan estos sistemas son de tipo factual (en las que se solicitan datos referidos a: persona, tiempo, localización, organización, medida, recuento, objeto, entre otros). Aplican técnicas cada vez más sofisticadas para extraer respuestas a partir del conjunto de documentos que constituye su base de datos. Un ejemplo, de entre los pocos disponibles en la Web, es *Start*, realizado por el *Massachusetts Institute of Technology* (*MIT*).

http://start.csail.mit.edu/

Las técnicas utilizadas son muy variadas y, según el enfoque adoptado por los diseñadores del sistema, van desde las estadísticas hasta el procesamiento del lenguaje natural –si bien suele ser ésta última la más frecuente–. Los diferentes criterios y decisiones que pueden tomarse en el diseño y elaboración de la arquitectura de un sistema de búsqueda de respuestas dan lugar a una heterogénea tipología de los mismos según sean las lenguas de trabajo, el contenido temático de los documentos de la base de datos, el nivel de organización de la información (estructurada o no) o el grado de interactividad con el usuario, entre otros aspectos. A continuación se indican los más frecuentemente identificados.

Idioma

El criterio de la lengua es muy popular en un gran número de las clasificaciones propuestas (**Adiwibowo**; **Adriani**, 2007; **Izquierdo** et al., 2007; **Roger** et al., 2007; **Solorio** et al., 2005). Se distingue entre sistemas monolingües y multilingües, aunque en algunos casos se consideran como un tercer tipo los multilingües que utilizan el inglés como idioma pivote (**García-Cumbreras** et al., 2006).

Tema

La cobertura temática de los documentos de la base de datos es tenida en cuenta por **Harabagiu** et al., 2000; **Magnini** et al., 2001; **Moldovan** et al., 2003; **Roger** et al., 2007. Así, mientras que unos son de dominio abierto, es decir, de tema general o multidisciplinar, no restringido, otros cuentan con una colección especializada en un tema. Como es de esperar, éstos son los que mejores resultados obtienen (**Wedgwood**, 2005).

Organización

El nivel de organización de la información contenida en los documentos indizados origina 3 tipos: los que trabajan con información estructurada, los que lo hacen con documentos semiestructurados (**Katz** et al., 2007) y los que gestionan una colección no estructurada (**Cucerzan**; **Agichtein**, 2005).

Interacción

La capacidad de interacción con el usuario es un aspecto que cobra aún mayor relevancia. Aunque es notable el interés creciente por los sistemas de búsqueda de respuestas capaces de mantener un diálogo continuado con los usuarios –y a éstos, a su vez, les resultan mucho más atractivos– la mayor parte de los sistemas aún no alcanzan un grado de comunicación conversacional. La ventaja de los sistemas plenamente interactivos es que utilizan cada nuevo *input* para conseguir cierta retroalimentación y afinar en la búsqueda de la respuesta más adecuada. Por ejemplo, están más preparados para resolver con éxito las preguntas enlazadas, que incluyen referencias a preguntas o a respuestas anteriores (¿quién descubrió la penicilina?, ¿dónde nació?). Los sistemas interactivos de búsqueda de respuestas encajan perfectamente con el enfoque de los tradicionales servicios de referencia. Las aplicaciones en este ámbito (**Pomerantz**, 2005) parten de la premisa de que la consulta primera del usuario suele ser errónea, ya que éste puede tener cierta dificultad para expresar su necesidad de información. Se cuestiona la validez de la pregunta inicial y se busca una reformulación más exacta ofreciendo mayor orientación durante la consulta. Por el

contrario, los sistemas de búsqueda de respuestas que no cuentan con la perspectiva del servicio de referencia, es decir, la inmensa mayoría, consideran que el usuario formula correctamente su pregunta. Establecen que hay una única respuesta posible –que puede estar contenida o no en el corpus de documentos– de modo que, disponiendo de la base de conocimientos adecuada, siempre se podrá satisfacer la necesidad de información.

Existen sistemas de búsqueda de respuestas que usan reconocimiento de voz para su implementación en dispositivos móviles (**Izquierdo** et al., 2007).

Arquitecturas

Un sistema de búsqueda de respuestas está constituido por módulos, cada uno de los cuales lleva a cabo una serie de tareas dentro de la arquitectura global del sistema. El número de ellos y las funciones que desempeñan varía según el enfoque desde el que se aborda el problema. **Pérez-Coutiño** et al. (2004) proponen cuatro módulos –uno para el procesamiento de la pregunta, otro de indexación, el de búsqueda y el dedicado a la selección de la respuesta–, mientras que otros autores consideran necesarios únicamente tres módulos –el destinado al análisis de la pregunta, el de recuperación de pasajes (los párrafos de los documentos donde se encuentra la respuesta) y el dedicado a la extracción de la respuesta propiamente dicha– (**Vicedo** et al., 2003).

"Los criterios y decisiones que se toman en el diseño de un sistema de búsqueda de respuestas dan lugar a tipos variados"

Lo más frecuente es establecer cuatro, que se implementan prácticamente del mismo modo en todos los sistemas (**Ittycheriah** et al., 2000; **Ko** et al., 2007). Estos módulos, que han de ser capaces de procesar el lenguaje natural de manera efectiva, corresponden a los cuatro procesos que lleva a cabo el sistema de búsqueda para la obtención de una respuesta:

– Análisis y clasificación del tipo de preguntas/respuestas. La primera tarea que realiza el sistema es identificar el tipo de pregunta planteada por el usuario. Su análisis implica su tratamiento lingüístico, y su clasificación se hace en función del tipo de respuesta que se espera obtener. La elaboración de un buen sistema de análisis y clasificación de las preguntas será clave para el éxito o fracaso del sistema.

– Recuperación de la información/expansión de la consulta. El procesamiento de la pregunta en la etapa anterior constituye el *input* para el motor de búsqueda, el cual devuelve documentos –o partes de documentos– relevantes de la base de datos.

– Reconocimiento de entidades nombradas. Ahora los documentos son divididos en pasajes –si bien esto puede hacerse en la fase anterior de recuperación– y se analizan mediante reconocedores de entidades con el fin de identificar los pasajes susceptibles de contener la respuesta (**Clarke** et al., 2000).

– Selección de la respuesta. Se selecciona de acuerdo con la clasificación de tipos de preguntas que se ha realizado y se identifica el pasaje que contiene la respuesta.

La clasificación de las preguntas es uno de los aspectos que más atención requieren, ya que un tercio de los errores se producen en esa etapa (**Moldovan** et al., 2003). En este sentido, el proceso de clasificación de las preguntas tiene dos objetivos (**Li**; **Roth**, 2004): por un lado, se reducen los tipos de respuestas esperadas. Por otro, se identifica el enfoque, estableciendo el tipo de pregunta en función del tipo de información que debe contener la respuesta esperada. Así por ejemplo, ante la pregunta "¿cuánto mide el monte Everest?", distinguiremos entre "monte Everest" (el objeto sobre el cual se solicita una información) y la altura del mismo (el enfoque desde el cual se solicita la información). Identificar tanto el objeto de la pregunta como el enfoque de la misma resulta una importante ayuda para localizar la respuesta adecuada.

Taxonomías para clasificar preguntas

La creación de un módulo para clasificar preguntas pasa por dos etapas fundamentales: 1) establecer la taxonomía de tipos de preguntas a las que, potencialmente, tendrá que enfrentarse el sistema y 2) diseñar el módulo clasificador para que identifique, etiquete y categorice cada pregunta del *input* en función de esa clasificación.

"Muchas de las herramientas y métodos existentes para tratar y clasificar las preguntas son para la lengua inglesa"

La taxonomía que se utilizará puede estar determinada por expertos o bien automáticamente (**Hacioglu**; **Ward**, 2003). Algunas propuestas para la clasificación de preguntas son fruto de una taxonomía elaborada manualmente a partir de una colección de preguntas y, al ser elaboradas por expertos suelen dar unos resultados muy positivos para colecciones *ad hoc,* si bien rara vez estos resultados son extrapolables al utilizar otra colección de preguntas y de documentos. Estas clasificaciones son te-

diosas de realizar y carecen de la flexibilidad necesaria para adaptarse a nuevos tipos de preguntas que puedan plantearse en el sistema. Por el contrario, las taxonomías elaboradas automáticamente surgen a partir de un conjunto lo suficientemente amplio de preguntas, aplicando diferentes tipos de técnicas. Independientemente de la técnica (manual o automática) elegida para desarrollar la taxonomía de tipos de preguntas, el siguiente paso consiste en preparar el módulo para que la utilice en la clasificación de las preguntas del usuario.

Las principales taxonomías que se han empleado para la clasificación de preguntas (**Pomerantz**, 2005) son las basadas en:

– "*Wh-words*". Se refiere a *who, what, when, where, why* y *how*. En español corresponden a las preguntas que comienzan por pronombres o adverbios interrogativos: *quién*, *qué*, *cuándo*, *dónde*, *por qué* y *cómo*. Es una taxonomía muy usada debido a su simplicidad (**Hovy** et al., 2001; **Moldovan** et al., 1999). Sin embargo no es muy efectiva para su uso en los sistemas de búsqueda de respuestas ya que no engloba todos los posibles tipos de preguntas, como por ejemplo "¿cuál es el principal producto de *Microsoft*?".

– Temas de preguntas. Sigue la filosofía bibliotecaria de organización por materias mediante el uso de herramientas como tesauros, encabezamientos de materias, etc. Puede resultar una clasificación más efectiva por la información que aporta a la pregunta, pero requiere el uso de fuentes externas o del reconocimiento de entidades nombradas, y supone una inversión de tiempo que actualmente no es recomendable utilizar como método para la clasificación de tipos de preguntas en sistemas de búsqueda de respuestas sino en el proceso de desambiguación de la pregunta, de ser necesario.

– Funciones de las respuestas esperadas. Según este criterio, la clasificación de las preguntas se hace teniendo en cuenta la función de la respuesta que se espera obtener. Es un método muy recomendable, pues añade información adicional al sistema. De este modo, a la pregunta "¿cuánto mide Pau Gasol?", el sistema buscaría un valor numérico y excluiría como respuestas candidatas a todas las que no incluyeran cifras, aumentando así la precisión de la consulta.

– Formas de las respuestas esperadas. Se trata de una taxonomía enfocada específicamente a los servicios de referencia y no es extensible a los sistemas de búsqueda de respuestas. Permite clasificar las preguntas en dos tipos principales: transacciones referenciales y transacciones direccionales. Así, por ejemplo, en una biblioteca, si el primer tipo se refiere a consultas que requieren la búsqueda de conocimiento (preguntas sobre un objeto, concepto, lugar, persona, entre otros), la consulta de fuentes y el análisis e interpretación de la pregunta, el segundo tipo engloba las preguntas relacionadas con los servicios bibliotecarios, como horarios de consulta, período de tiempo en préstamo, etc.

– Tipos de fuentes según el tipo de respuesta que se espera. El enfoque para la clasificación de las preguntas recae sobre la función de la respuesta pero en este caso se trata de identificar la fuente de información en la que habría que lanzar la búsqueda. De tratarse de un hecho histórico, por ejemplo, el sistema habría de extraer la respuesta de una enciclopedia; si fuera una pregunta de definición, iría directamente a un diccionario, y de tratarse de una pregunta sobre un suceso reciente lanzaría la consulta en periódicos y agencias de noticias. Esta taxonomía podría utilizarse de apoyo, de cara a la desambigüación o la expansión de la consulta.

> "Un tercio de los errores se producen en la etapa de clasificación de las preguntas"

Esos criterios de clasificación no son excluyentes sino que pueden combinarse para lograr un mejor procesamiento de las preguntas. De hecho, lo más habitual es que se combinen los tres primeros criterios (**Pomerantz**, 2005) de manera secuencial, por lo que progresivamente el sistema va especificando el tipo de pregunta formulada. De este modo un primer filtrado correspondería a la clasificación mediante *wh-words*". Si se formula una pregunta como "¿quién es el primer ministro del Reino Unido?", rápidamente se detectaría el tipo de información que se solicita (una persona) y mediante un reconocedor de entidades nombradas se podría localizar la respuesta. No obstante, habría una gran cantidad de preguntas que se escaparían a esta primera taxonomía, incluso preguntas que aun perteneciendo a alguna de sus categorías hayan sido reformuladas y no utilicen las partículas interrogativas que definen el tipo de pregunta. Por ejemplo, la pregunta "¿a qué partido político pertenece Gordon Brown?" no tendría cabida. Sin embargo, aplicando la taxonomía desarrollada a partir del criterio temático de clasificación de preguntas, el sistema sí podría identificar la respuesta correcta. Combinando ambos criterios se habría resuelto la clasificación de una gran cantidad de tipos de preguntas, pero el margen de error sería, aún así, demasiado amplio. Por ello, se podría hacer una tercera y definitiva aproximación según las funciones de las respuestas esperadas. Así se conseguiría más precisión en preguntas como "¿en qué año fue nombrado Gordon Brown primer ministro del Reino Unido?", ya que se evitarían posibles respuestas erróneas o se res-

tringiría la longitud de la respuesta al valor numérico correspondiente al año.

Por último mencionamos una de las propuestas de más éxito, la taxonomía de **Li** y **Roth** (2004), la cual presenta una estructura jerárquica de hasta 50 categorías diferentes, englobadas en seis clases principales:

– Abreviatura:

Abreviatura, abreviatura expandida.

– Entidad:

Acontecimiento, animal, color, creativa, cuerpo, comida, deporte, enfermedad, idioma, instrumento, moneda, otros, palabra, planta, producto, religión, símbolo, sustancia, técnica, término, vehículo.

– Descripción:

Definición, clase, descripción, manera, razón.

– Humano:

Grupo, individuo, título, descripción.

– Localización:

Ciudad, montaña, otros, país, estado.

– Numérico:

Código, cuenta, fecha, distancia, dinero, orden, otros, período, porcentaje, velocidad, tiempo, tamaño, peso.

Tratamiento lingüístico de las preguntas

Uno de los principales inconvenientes que presentan estos sistemas es que están ostensiblemente limitados por el idioma para el que fueron diseñados. A pesar de que se está investigando en la búsqueda de métodos independientes de la lengua (**Solorio** et al., 2004; **Whittaker**; **Furui**, 2006), en la práctica se sigue trabajando en los de una sola. Es en las conferencias *CLEF* (*Cross-language evaluation forum*) donde se dedica especial atención a los sistemas multilingües de búsqueda de respuestas, y desde 2003 disponen de una sección dedicada exclusivamente a esta tarea (**Braschler**; **Peters**, 2004).

Como se muestra en la figura 1, antes de proceder a la clasificación de las preguntas el sistema las tendrá que someter a un proceso de análisis y tratamiento con el fin de eliminar las posibles ambigüedades que puedan presentar y prepararlas para su categorización.

Al igual que para los demás aspectos, muchas de las herramientas y de las metodologías desarrolladas para llevar a cabo el tratamiento de las preguntas y su clasificación, se centran en la lengua inglesa (**Solorio** et al., 2004) por lo que cuando quieren utilizarse en sistemas que trabajan con la lengua española es necesario adaptarlas o bien crear otras nuevas que se adecuen a las singularidades del español.

Liddy (1998, 2003) señala siete niveles lingüísticos en los que se ha de trabajar en el PLN: fonético, morfológico, léxico, sintáctico, semántico, discursivo y pragmático.

Este planteamiento es igualmente aplicable al tratamiento de las preguntas que constituyen el *input* en los sistemas de búsqueda de respuestas, si bien dependiendo de las características del sistema se deberá incidir especialmente en unos niveles o en otros. En este sentido, **Pomerantz** (2005) indica que el análisis lingüístico debe centrarse en los cuatro niveles superiores –sintáctico, semántico, discursivo y pragmático–. Podemos añadir el análisis fonético para los sistemas que trabajen con voz. En cada caso habría que aplicar diferentes técnicas y herramientas de procesamiento de lenguaje natural.

A continuación se indican cuáles son las distintas etapas de análisis a las que ha de someterse el texto de las preguntas en los sistemas de búsqueda de respuestas. Como no siempre resulta posible implantar soluciones *ad hoc*, y a modo de orientación, mencionamos algunas herramientas disponibles que pueden ayudar en ciertas tareas.

– Análisis fonético. Es el nivel de análisis lingüístico más básico. Pretende interpretar los sonidos y hacerlos inteligibles para el sistema. Se usa en sistemas que aplican reconocimiento y procesamiento del habla, como el anteriormente mencionado.

– Análisis morfológico. En esta etapa se analizan los componentes que conforman cada palabra. Se extraen los prefijos, sufijos, rasgos flexivos y raíces, entre otros elementos. Aquí se incluye, por ejemplo, la aplicación de las técnicas de *stemming* (o reducción de una palabra a su raíz) que facilitan el tratamiento del texto de la pregunta y permiten la expansión de la consulta. Para el *stemming* podría utilizarse una versión del *algoritmo de Porter* (**Porter**, 1980) adaptado al español, tal como *Snowball*[1] que resolverían algunos problemas de ambigüedad. El análisis morfológico se hace necesario ya que, a pesar de encontrarnos aún en una de las primeras etapas de aproximación lingüística al texto del *input*, conviene tratarlo adecuadamente para facilitar la extracción del significado de las palabras y el significado de la pregunta en las etapas posteriores.

http://snowball.tartarus.org/

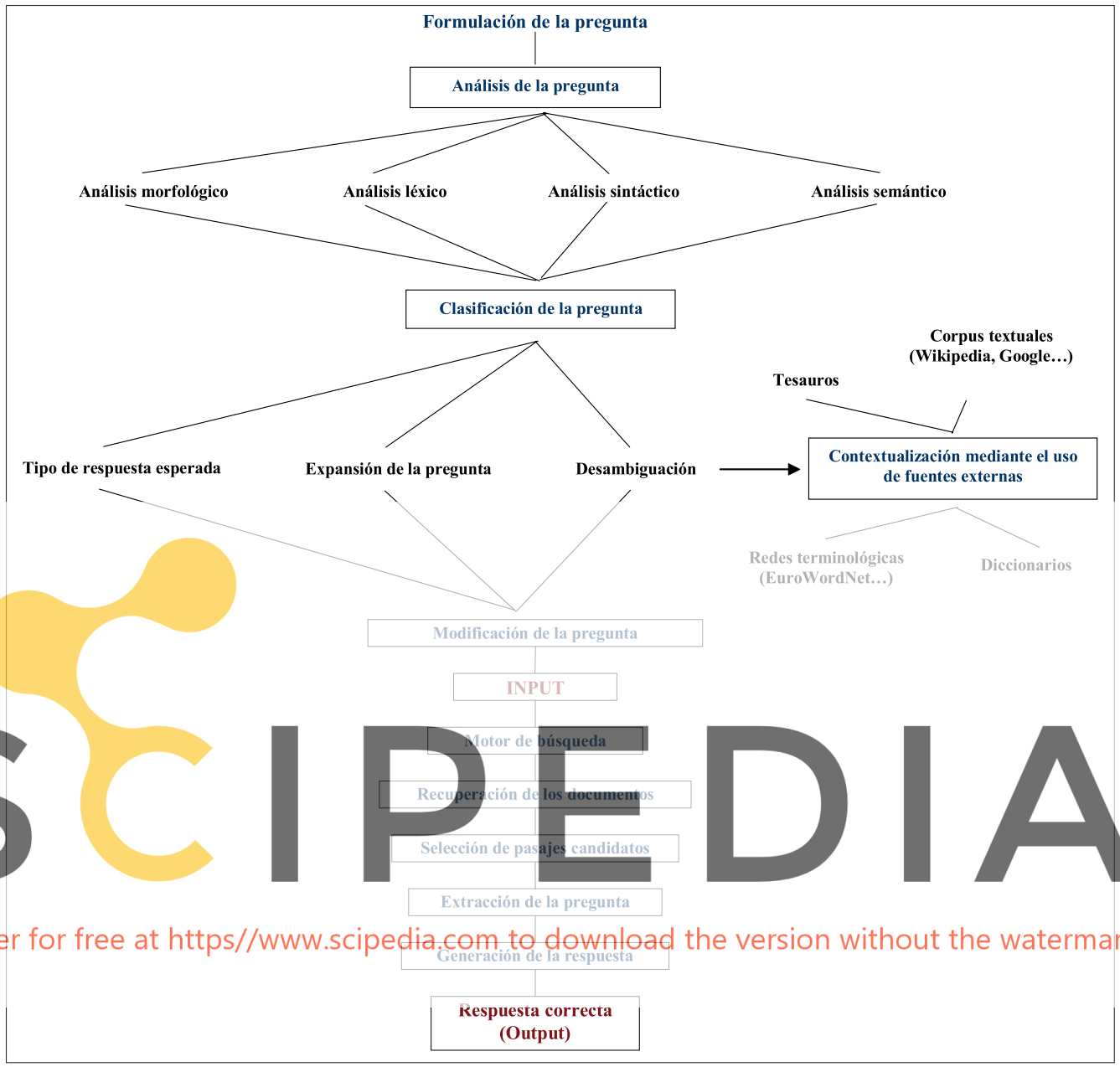

Figura 1. Funcionamiento de un sistema de búsqueda de respuestas basado en el PLN

– Análisis léxico. Se realiza un análisis de las palabras desde una perspectiva intralingüística para desentrañar el significado léxico, es decir, se definen y delimitan las unidades léxicas que componen la pregunta, pudiéndose hacer uso de un analizador léxico. Con ayuda del mismo se filtran los elementos que no aportan significado a la pregunta como, por ejemplo, las palabras vacías y los signos ortográficos y de puntuación. De este modo nos quedamos únicamente con las palabras realmente relevantes.

– Análisis sintáctico. En este caso la unidad de estudio no es la palabra sino la frase completa, analizando su estructura gramatical y determinando la información significativa para la selección de la respuesta. Por ejemplo, ante una pregunta del tipo "¿dónde vive José Luís Rodríguez Zapatero?", se deberá identificar José Luís Rodríguez Zapatero como la entidad sobre la que se demanda información, cosa que se hace analizando sintácticamente cuál es el sujeto de la oración. Ante la pregunta "¿en qué posición juega Sergio Ramos en el Real Madrid?", vemos que nos encontramos ante dos entidades nominativas: el jugador Sergio Ramos y el equipo de fútbol Real Madrid. No obstante, la pregunta se centra exclusivamente en el jugador y no en el equipo. De hecho Real Madrid sería una información irrelevante que podría confundir al sistema al tratarse de una entidad fácilmente reconocible; sin embargo, al ser reconocida como un complemento, quedaría en un segundo plano. Hay varios analizadores sintácticos desarrollados para el español (**Bick**, 2006; **Carreras** et al., 2004) y muchos de ellos están disponibles en internet de forma gratuita[2].

– Análisis semántico. En este nivel de análisis lingüístico se barajan los posibles significados de cada

	Niveles lingüísticos
-	fonético
	morfológico
	léxico
Profundidad de análisis	sintáctico
	semántico
	discursivo
+	pragmático

*Figura 2. Profundidad del análisis para el PLN en función de los niveles lingüísticos (**Liddy**, 1998, 2003)*

frase y se busca desambiguar la acepción de las palabras por el contexto. Será especialmente relevante el tratamiento del lenguaje para las palabras polisémicas, en cuyo caso se deberá intentar identificar el significado correcto a través del contexto en el que esté la palabra.

Sin embargo, las preguntas factuales planteadas por los usuarios en los sistemas de búsqueda de respuestas generalmente carecen de contexto, por lo que entonces la desambiguación de las palabras puede resolverse mediante el uso de fuentes externas, como la Web (**Adiwibowo**; **Adriani**, 2007). Esta solución tiene sus inconvenientes, pero ayuda a analizar semánticamente la pregunta para desambiguarla. Y, por supuesto, supone proporcionar un valor añadido a la pregunta del usuario que, como texto aislado y breve que es, resulta bastante difícil contextualizar.

Otra forma de aportar información a las preguntas es mediante el uso de reconocedores de entidades nombradas (REN), los cuales favorecen la identificación de ciertos objetos en los textos. Por entidades nombradas se entienden los nombres de personas, organizaciones, localizaciones, fechas o cantidades (**Balbontín**; **Sánchez**, 2004). Un REN es una herramienta capaz de identificarlos, favoreciendo de este modo la comprensión del sistema de cara al análisis y tratamiento de las preguntas.

– Análisis discursivo. Está enfocado al discurso, e interpreta la estructura y el significado de los textos compuestos por más de una frase, siendo adecuado para aplicar técnicas de PLN a textos largos. Como las preguntas en los sistemas de búsqueda de respuestas suelen ser oraciones breves, no es necesario aplicarlo aquí.

– Análisis pragmático. En este nivel de análisis lingüístico se pretende entender el propósito de la lengua en situaciones concretas, particularmente los aspectos del lenguaje que requieren de un conocimiento general. Se pretende desambiguar los elementos de la pregunta que, aun habiendo identificado claramente su significado semántico, no acaban de ser bien interpretados por el sistema al ser necesario un cierto acervo de conocimientos para poder ser entendidos completamente. La pregunta "¿cuál fue el primer disco de los cuatro de Liverpool?" como podemos suponer, se refiere a Los Beatles y *a priori* debería ser relativamente fácil identificar la respuesta. Pero el hecho de utilizar un apodo para referirse al grupo musical complica la capacidad del sistema. Esta circunstancia podría subsanarse, al menos en parte, con el uso de fuentes externas: enciclopedias como *Wikipedia* o de buscadores como *Google*, que permiten discernir personajes y sucesos que son de conocimiento general y despejar rápidamente esas lagunas.

Conclusiones y líneas futuras

A pesar de su corta trayectoria, los sistemas de búsqueda de respuestas constituyen una interesante opción de cara a la recuperación de información y a la satisfacción de las necesidades de los usuarios. Permiten una mayor usabilidad y una mayor interacción con el usuario mejorando su experiencia a la hora de enfrentarse a la búsqueda de información. No obstante, son muchos los obstáculos que se presentan en la actualidad, sobre todo aquellos relacionados con el procesamiento del lenguaje natural. Iniciativas como las de las conferencias *TREC* y *CLEF* intentan aunar esfuerzos y constituirse en referentes para los investigadores que trabajan en estos temas.

Los futuros estudios que pretendan contribuir a resolver algunos de los muchos problemas que estos sistemas aún presentan, deberían centrarse en los siguientes aspectos (**Burger** et al., 2001):

– taxonomías de tipos de preguntas;
– procesamiento de las preguntas;
– establecer el contexto de las preguntas;
– fuentes de información;
– extracción de las respuestas;
– formulación de la respuesta;
– respuesta en tiempo real;
– sistemas multilingües;
– sistemas interactivos;
– razonamiento avanzado;
– captación de tipos de usuarios; y
– sistemas colaborativos.

Como se puede comprobar, se abordan tanto temas de procesos que deben llevar a cabo estos sistemas, como tipos de sistemas y distintos enfoques para implementarlos. Es importante destacar la necesidad de trabajar en sistemas multilingües y en sistemas monolingües que hagan uso del español.

Notas

1. *Snowball* tiene versiones de hasta 15 idiomas. *http://snowball.tartarus.org/*

2. *VISL* es un analizador sintáctico disponible para varios idiomas. Su ver-

sión en español se llama *Hispal*. *FreeLing* es un analizador sintáctico open source de descarga gratuita. *Stilus* se puede probar por internet.
http://beta.visl.sdu.dk/
http://garraf.epsevg.upc.es/freeling/
http://www.connexor.eu/

Bibliografía

Adiwibowo, S.; **Adriani, M.** "Finding answers using resources in the internet". En: *Working notes for the CLEF 2007 Workshop.* *http://www.clef-campaign.org/2007/working_notes/adiwibowoCLEF2007.pdf*

Balbontín-Gutiérrez, A.; **Sánchez-Martín, J. J.** "Spner – Reconocedor de entidades nombradas para el español". En: *Cuarta conferencia de procesamiento del lenguaje natural de la Universidad Europea de Madrid*, 2004. *http://www.esp.uem.es/~jmgomez/plenum/plenum4/04.pdf*

Berners-Lee, T.; **Hendler, J.**; **Lassila, O.** "The semantic web". *Scientific American*, 2001, v. 284, n. 5, pp. 34-43.

Bick, E. "A constraint grammar-based parser for Spanish". En: *Proceedings of TIL 2006 - 4th Workshop on information and human language technology*, 2006.

Braschler, M.; **Peters, C.** "Cross-language evaluation forum: objectives, results, achievements". *Information retrieval*, 2004, v. 7, pp. 7-31.

Burger, J.; **Cardie, C.**; **Chaudhri, V.**; **Gaizauskas, R.** et al. "Issues, tasks and program structures to roadmap research in question & answering". *National Institute of Standards and Technology.* http://www.icst.pku.edu.cn/course/TextMining/07-08Spring/%E5%8F%82%E8%80%83%E6%96%87%E7%8C%AE/11-01 Issues, Tasks and Program Structures to Roadmap Research in Question & Answering.pdf

Buscaldi, D. et al. "The *UPV* at QA@CLEF 2007". *8th Intl. cross-language evaluation forum CLEF-2007 working notes,* 2007.

Carreras, X.; **Chao, I.**; **Padró, L.**; **Padró, M.** "FreeLing: an open-source suite of language analyzers". En: *Proceedings of the 4th Intl. conf. on language resources and evaluation (LREC)*, 2004, pp. 239-242. *http://www.lsi.upc.edu/~nlp/papers/carreras04.pdf*

Clarke, C. L. A. et al. "Question answering by passage selection". En: *The 9th text retrieval conference (TREC 9)*, 2000.

Cucerzan, S.; **Agichtein, E.** "Factoid question answering over unstructured and structured web content". En: *ACM Conference on information and knowledge management,* 2005.

Durme, B. Van et al. "Towards light semantic processing for question answering". En: *Proceedings of human language technology conference (HLT-Naacl),* 2003.

García-Cumbreras, M. A.; **Ureña-López, L. A.**; **Martínez-Santiago, F.** "Bruja: question classification for Spanish. Using machine translation and an English classifier". En: *EACL 2006 Workshop on multilingual question answering – MLQA06*, 2006, pp. 39-44.

Hacioglu, K.; **Ward, W.** "Question classification with support vector machines and error correcting codes". En: *Proceedings of the human language technology conference (HLT-Naacl)*, 2003, pp. 28-30.

Harabagiu, S. M.; **Pasca, M. A.**; **Maiorano, S. J.** "Experiments with open-domain textual question answering". En: *Proceedings of the Coling-2000. Association for Computational Linguistics / Morgan Kaufmann*, 2000.

Hovy, E.; **Hermjakob, U.**; **Lin, C.-Y.** "The use of external knowledge in factoid QA". En: *Tenth text retrieval conference (TREC 10)*, 2001.

Ittycheriah, A. et al. "IBM's statistical question answering system". En: *Proceedings of the Nineth text retrieval conference* (TREC 9), 2000.

Izquierdo, R.; **Ferrández, O.**; **Ferrández, S.**; **Vicedo, J. L.**; **Martínez, P.**; **Suárez, A.** "QALL-ME: question answering learning technologies in a multiLingual and multiModal environment". *Revista procesamiento del lenguaje natural*, 2007, n. 38, pp. 43-47. *http://www.sepln.org/revistaSEPLN/revista/38/SEPLN38.pdf*

Katz, B. et al. "Answering English questions using foreign-language, semi-structured sources". En: *Proceedings of the first IEEE intl. conf. on semantic computing (ICSC 2007)*, 2007, pp. 439-445.

Ko, J.; **Mitamura, T.**; **Nyberg, E.** "Language-independent probabilistic answer ranking for question answering". En: *Proceedings of the 45th annual meeting of the Association of Computational Linguistics (ACL 07)*, 2007, pp. 784-791.

Li, X.; **Roth, D.** "Learning question classifiers". En: *Proceedings of the 19th Intl. conf. on computational linguistics (Coling 02)*, 2004.

Liddy, E. D. "Enhanced text retrieval using natural processing". *Bulletin of the American Society for Information Science*, 1998, v. 24, n. 4, pp. 14-16.

Liddy, E. D. "Natural language processing". En: *Encyclopedia of library and information science*. 2ª ed. NY: Marcel Decker Inc., 2003

Magnini, B.; **Negri, M.**; **Prevete, R.** "Open domain question/answering on the Web". En: *Congress of the Italian Association for Artificial Intelligence*, 2001, n. 2175, pp. 273-284.

Moldovan, D. et al. "Lasso: A tool for surfing the answer net". En: *Eighth text retrieval conference (TREC 8),* 1999.

Moldovan, D. et al. "Performance issues and error analysis in an open-domain question answering system". *ACM transactions on information systems*, 2003, v. 21, n. 2, pp. 133-154.

Pérez-Coutiño, M.; **Solorio, T.**; **Montes-Gómez, M.**; **López-López, A.**; **Villaseñor-Pineda, L.** "The use of lexical context in question answering for Spanish". En: *Workshop of the cross language evaluation forum (CLEF 2004)*. Bath, UK. September 2004. *http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/46.pdf*

Pomerantz, J. "A linguistic analysis of question taxonomies". *Journal of the American Society for Information Science and Technology*, 2005, v. 56, n. 7, pp. 715-728.

Porter, M. F. "An algorithm for suffix stripping". *Program*, 1980, v. 14, n. 3, pp. 130-137.

Roger, S.; **Ferrández, A.**; **Peral, J.**; **Ferrández, S.**; **López-Moreno, P.** "An inference mechanism for question answering". *Journal of computer science & technology*, 2007, n. 1, pp. 21-27. *http://journal.info.unlp.edu.ar/Journal/journal19/papers/JCST-Mar07-4.pdf*

Solorio, T.; **Pérez-Coutiño, M.**; **Montes-Gómez, M.**; **Villaseñor-Pineda, L.**; **López-López, A.** "A language independent method for question classification". En: *The 20th Intl. conf. on computer linguistics, Coling 04*, 2004, v. 2, pp. 1374-1380.

Solorio, T.; **Pérez-Coutiño, M.**; **Montes-Gómez, M.**; **Villaseñor-Pineda, L.**; **López-López, A.** "Question classification in Spanish and Portuguese". En: *Conference on intelligent text processing and computational linguistics, CICLing 2005. Lecture notes in computer science*, 2005, v. 3406, pp. 612-619.

Vállez, M. "Web semántica y procesamiento del lenguaje natural" En: **Codina, L.**; **Rovira, C.**; **Marcos, M. C.** (eds.) Web semántica y sistemas de información documental. TREA, 2007. *http://www.semanticaweb.net/libro/autores.htm*

Vicedo, J. L.; **Izquierdo, R.**; **Muñoz, R.** "Question answering in Spanish". En: *Proceedings of CLEF 2003*, 2003, pp. 541-548.

Wedgwood, J. A. "Medical question-answering framework". En: *AMIA 2005 Symposium proceedings*, p. 1.150.

Whittaker, E.; **Furui, S.** "Progress on a language independent approach to question answering". En: *Proceedings symposium on large-scale knowledge resources* 2006, v. 3, pp. 171-174.

María-Dolores Olvera-Lobo *y* ***Nicolás Robinson-García****, Universidad de Granada, Facultad de Comunicación y Documentación, Campus Cartuja, 18007 Granada.*
Tel.: +34-958 243 478
molvera@ugr.es
elrobinster@gmail.com